**MODEL KÜÇÜLTME TEKNİKLERİ**

### **(Knowledge Distillation, Parameter Sharing, Quantization ve Matrix Factorization)**

Son zamanlarda derin öğrenme modelleri büyük gelişmeler kaydetti özellikle görüntü işleme, doğal dil işleme ses tanıma…gibi. Ancak bu modellerin yüksek hesaplama maliyetlerinden ve büyük boyutlu olmasına kadar birçok etkenden ötürü düşük kaynaklı ortamlarda (mobil cihazlar, IoT cihazları gibi)kullanımlarını zorlaştırmaktadır. Bu soruna çözüm olarak model küçültme teknikleri geliştirilmiştir. Bu yaptığım literatür taramasında ise Knowladge Distillation, Parameter Sharing, Quantization ve Matrix Factorization gibi popüler model küçültme teknikleri incelenecektir.

1. **Knowledge Distillation (Bilgi Damıtma);**

Bu teknikte büyük modele öğretmen model küçük modele öğrenci model diyeceğiz. Büyük bir modelin (öğretmen model) bilgisini daha küçük bir modele (öğrenci model) aktarmak için kullanılan bir tekniktir. Bu yöntem özellikle derin sinir ağlarının boyutunu küçültmek için etkilidir.

* **Temel Çalışma Prensibi**: Öğretmen model, eğitim verileri üzerinde eğitilir ve bu modelin çıktıları (soft targets) öğrenci model için bir rehber olarak kullanılır. Öğrenci model, öğretmen modelin çıktılarını taklit ederek eğitilir. Bu sayede, öğrenci model, öğretmen modelin performansına yakın bir performans sergileyebilir (Hinton et al., 2015).
* **Avantajları**:
  + Öğrenci model, öğretmen modelden daha az parametreye sahiptir.
  + Öğrenci model, öğretmen modelin genelleme yeteneğini miras alır.
  + Düşük kaynaklı cihazlarda çalıştırılabilir.
* **Literatürdeki Örnekler**:
  + Hinton et al. (2015), "Distilling the Knowledge in a Neural Network" makalesinde bu tekniği tanıtmış ve popüler hale getirmiştir.
  + Bu yöntem, özellikle görüntü sınıflandırma ve doğal dil işleme gibi alanlarda yaygın olarak kullanılmaktadır (Gou et al., 2021).

1. **Parameter Sharing (Parametre Paylaşımı);**

Parametre paylaşım modeli. Farklı katmanlar ve bileşenleri arasında aynı parametrelerin kullanılmasını içeren bir tekniktir. Tekrarlayan sinir ağları (RNN) ve transformatör modellerinde yaygın olarak kullanılan bir yöntemdir.

**2.1. Temel Çalışma Prensibi**

Modelin farklı kısımları arasında aynı ağırlıklar paylaşılır. Örneğin, RNN'lerde her zaman adımında aynı ağırlık matrisi kullanılır. Bu, modelin parametre sayısını önemli ölçüde azaltır (Vaswani et al., 2017).

#### **2.2. Avantajları**

* **Model Boyutunda Küçülme**: Model boyutu küçülür.
* **Eğitim Süresi ve Bellek Kullanımında Azalma**: Eğitim süresi ve bellek kullanımı azalır.
* **Dil Modellerinde Etkililik**: Özellikle dil modellerinde etkilidir.

#### **2.3. Literatürdeki Örnekler**

* Transformers modellerinde (Vaswani et al., 2017), parametre paylaşımı kullanılarak model boyutu küçültülmüştür.
* ALBERT (A Lite BERT) modeli, parametre paylaşımı kullanarak BERT modelinin boyutunu küçültmüştür (Lan et al., 2019).
* Özellikle, ALBERT modeli, parametre paylaşımı sayesinde BERT'e kıyasla %90'a varan parametre azaltımı sağlamıştır.

Kuantizasyon, modelin ağırlıklarını ve aktivasyonlarını daha düşük hassasiyetli sayısal değerlere (örneğin, 32-bit yerine 8-bit) dönüştürme işlemidir. Bu yöntem, modelin bellek kullanımını ve hesaplama maliyetini azaltır

**3. Quantization (Kuantizasyon)**

#### **3.1. Temel Çalışma Prensibi**

Modelin ağırlıkları ve aktivasyonları, daha düşük bit genişliğine sahip değerlere dönüştürülür. Bu işlem, modelin performansını korurken bellek ve hesaplama gereksinimlerini azaltır (Jacob et al., 2018).

#### **3.2. Avantajları**

* **Model Boyutunda Küçülme**: Model boyutu önemli ölçüde küçülür.
* **Hesaplama Hızında Artış**: Hesaplama hızı artar.
* **Enerji Verimliliği**: Enerji verimliliği artar.

#### **3.3. Literatürdeki Örnekler**

* Jacob et al. (2018), "Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference" makalesinde kuantizasyon tekniklerini detaylı bir şekilde incelemiştir.
* TensorFlow Lite ve PyTorch gibi popüler derin öğrenme kütüphaneleri, kuantizasyonu desteklemektedir.
* Özellikle, mobil cihazlarda çalıştırılan modellerde kuantizasyon yaygın olarak kullanılmaktadır (Krishnamoorthi, 2018).

### **Matrix Factorization (Matris Faktorizasyonu)**

Matris faktorizasyonu, büyük matrislerin daha küçük matrislerin çarpımı olarak ifade edilmesini sağlayan bir tekniktir. Bu yöntem, özellikle modelin ağırlık matrislerini küçültmek için kullanılır.

#### **4.1. Temel Çalışma Prensibi**

Büyük bir ağırlık matrisi, iki veya daha fazla küçük matrisin çarpımı olarak ifade edilir. Bu sayede, modelin parametre sayısı azaltılır (Sainath et al., 2013).

#### **4.2. Avantajları**

* **Model Boyutunda Küçülme**: Model boyutu küçülür.
* **Hesaplama Karmaşıklığında Azalma**: Hesaplama karmaşıklığı azalır.
* **Doğrusal Cebir Tabanlı Modellerde Etkililik**: Özellikle doğrusal cebir tabanlı modellerde etkilidir.

#### **4.3. Literatürdeki Örnekler**

* SVD (Singular Value Decomposition) gibi teknikler, matris faktorizasyonu için yaygın olarak kullanılır.
* Low-Rank Factorization teknikleri, derin öğrenme modellerinde ağırlık matrislerini küçültmek için kullanılmıştır (Sainath et al., 2013).
* Özellikle, konuşma tanıma modellerinde matris faktorizasyonu, model boyutunu küçültmek için başarıyla uygulanmıştır.

### **Sonuç**

Model küçültme teknikleri, derin öğrenme modellerinin boyutunu ve hesaplama maliyetini azaltmak için kritik bir rol oynar. **Knowledge Distillation**, **Parameter Sharing**, **Quantization** ve **Matrix Factorization**gibi yöntemler, modellerin düşük kaynaklı ortamlarda çalıştırılabilmesini sağlar. Bu teknikler, özellikle mobil cihazlar, IoT cihazları ve gerçek zamanlı uygulamalar için büyük önem taşır. Literatürde bu tekniklerin farklı varyasyonları ve uygulamaları bulunmaktadır, ancak temel prensipler ve avantajlar benzerdir. Gelecekte, bu tekniklerin daha da geliştirilerek daha verimli ve hafif modellerin oluşturulması beklenmektedir.

### **Kaynakça**

1. Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the Knowledge in a Neural Network. arXiv preprint arXiv:1503.02531.
2. Vaswani, A., et al. (2017). Attention is All You Need. Advances in Neural Information Processing Systems, 30.
3. Jacob, B., et al. (2018). Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
4. Lan, Z., et al. (2019). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. arXiv preprint arXiv:1909.11942.
5. Gou, J., et al. (2021). Knowledge Distillation: A Survey. International Journal of Computer Vision, 129(6), 1789-1819.
6. Sainath, T. N., et al. (2013). Low-rank matrix factorization for deep neural network training with high-dimensional output targets. IEEE International Conference on Acoustics, Speech and Signal Processing.
7. Tang, R., et al. (2019). Distilling Task-Specific Knowledge from BERT into Simple Neural Networks. arXiv preprint arXiv:1903.12136.
8. Krishnamoorthi, R. (2018). Quantizing deep convolutional networks for efficient inference: A whitepaper. arXiv preprint arXiv:1806.08342.

Berat Zeki Nihan

\*\*Not: Sayın Hocam Son gün haberim olduğum için daha iyi işler çıkarabilecekken daha basit kaldım.

Saygılarımla.